2012

L1pred: A Sequence-Based Prediction Tool for Catalytic Residues in Enzymes with the L1-logreg Classifier

Yongchao Dou
University of Nebraska-Lincoln

Jun Wang
Shanghai Normal University

Jialiang Yang
Chinese Academy of Sciences

Chi Zhang
University of Nebraska-Lincoln, zhang.chi@unl.edu

OPEN ACCESS Freely available online

PLoS one

L1pred: A Sequence-Based Prediction Tool for Catalytic Residues in Enzymes with the L1-logreg Classifier

Yongchao Dou[1], Jun Wang[2,3], Jialiang Yang[4], Chi Zhang[1]*

1 School of Biological Sciences, Center for Plant Science and Innovation, University of Nebraska, Lincoln, Nebraska, United States of America, **2** Scientific Computing Key Laboratory of Shanghai Universities, Shanghai, People's Republic of China, **3** Department of Mathematics, Shanghai Normal University, Shanghai, People's Republic of China, **4** MPI-Institute of Computational Biology, Chinese Academy of Sciences, Shanghai, People's Republic of China

Abstract

To understand enzyme functions, identifying the catalytic residues is a usual first step. Moreover, knowledge about catalytic residues is also useful for protein engineering and drug-design. However, to experimentally identify catalytic residues remains challenging for reasons of time and cost. Therefore, computational methods have been explored to predict catalytic residues. Here, we developed a new algorithm, L1pred, for catalytic residue prediction, by using the L1-logreg classifier to integrate eight sequence-based scoring functions. We tested L1pred and compared it against several existing sequence-based methods on carefully designed datasets Data604 and Data63. With ten-fold cross-validation, L1pred showed the area under precision-recall curve (AUPR) and the area under ROC curve (AUC) of 0.2198 and 0.9494 on the training dataset, Data604, respectively. In addition, on the independent test dataset, Data63, it showed the AUPR and AUC values of 0.2636 and 0.9375, respectively. Compared with other sequence-based methods, L1pred showed the best performance on both datasets. We also analyzed the importance of each attribute in the algorithm, and found that all the scores contributed more or less equally to the L1pred performance.

Citation: Dou Y, Wang J, Yang J, Zhang C (2012) L1pred: A Sequence-Based Prediction Tool for Catalytic Residues in Enzymes with the L1-logreg Classifier. PLoS ONE 7(4): e35666. doi:10.1371/journal.pone.0035666

Editor: Iddo Friedberg, Miami University, United States of America

Received January 4, 2012; **Accepted** March 19, 2012; **Published** April 27, 2012

Funding: This work was partially supported by the National Natural Science Foundation of China (No. 10731040), Shanghai Leading Academic Discipline Project (No. S30405) and Innovation Program of Shanghai Municipal Education Commission (No. 09zz134). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. No additional external funding received for this study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: czhang5@unl.edu

Introduction

Enzymes are very important because they act as catalysts for almost all chemical reactions in a cell to make the reaction rates sufficient for life. Identifying catalytic residues of enzymes is a crucial step towards understanding their functions. The knowledge on catalytic residues can further help design novel proteins with new functions and hence be useful for drug-design. Despite the importance, the number of proteins with known catalytic sites compared with the huge number of enzymes is still small, as it is often expensive and time consuming to experimentally identify catalytic residues. Fortunately, computational methods have become an important tool to predict catalytic residues with more and more annotated enzymes available.

In the past decade and a half, many computational methods have been developed to predict catalytic residues on given enzymes. The forerunners only considered protein sequence conservation information [1–12]. Prediction methods were then improved by incorporating phylogenetic motifs [13,14], phylogenetic trees [15,16], predicted structural information [17], and amino acids stereo-chemical properties [18–20] with conservation information. With increasing number of solved protein structures, structural information was also taken into account by many algorithms, however, which were limited only to proteins with known structures [21–31]. Meanwhile, Brylinski *et al.* developed a method to recognize protein active sites based on the analysis of hydrophobicity distribution in protein molecules [32]. In recent years, machine learning algorithms, such as Support Vector Machine-based (SVM) and Neural Network-based (NN), were used to develop new catalytic residue prediction methods [33–40]. The machine-learning algorithms can easily integrate various chemical and physical features of residues, such as sequence conservation, residue types, cumulative hydrophobicity, secondary structure, and relative solvent accessibility. For instance, Gutteridge *et al.* [33] used NN to incorporate six attributes extracted from both protein sequences and structures. Petrova and Wu [35] developed a similar method but using SVM. Zhang *et al.* [37] proposed an SVM-based method, called CRpred, which used sequence-derived attributes only. Youn reviewed several frequently used features and ranked their performance based on their ability to distinguish catalytic residues from non-catalytic ones; the top-ranked features are sequence conservation, structural conservation, uniqueness of a residue's structural environment, solvent accessibility, and residue hydrophobicity [36]. The flourishing efforts demonstrated promising potentials of computational methods on this research front, yet higher prediction accuracy is still needed for better performance.

In this manuscript, we developed a tool to predict enzyme catalytic residues. This tool is called L1pred because it uses the L1-logreg classifier, which is an implementation of the interior-point method for L1-regularized logistic regression [41]. Eight scoring functions used by L1pred to abstract protein sequence chemical/physical characteristics are residue type (RT), overlapping properties (OP), averaged cumulative hydrophobicity (ACH),

predicted protein secondary structure (SS), predicted accessible surface area (ASA), Jensen-Shannon divergence (JSD) conservation score, the combination of relative entropy of Venn diagram and JSD conservation score (VJSD), and Consurf score. We compared our method with others, such as JSD [5], VJSD [19], Consurf [42] and CRpred [37], and L1pred was shown to have the highest AUPR and AUC value for the same datasets. The curated datasets, the trained model, and the source code files are available at http://sysbio.unl.edu/L1pred.

Results

Results on the Dataset Data604

The parameters of L1pred were trained on the dataset Data604. The performance of L1pred achieved the optimal point at window-size = 6 and $\lambda = 0.002$; the corresponding maximal AUPR and AUC are 0.2198 and 0.9494, respectively. In the rest of the study, we applied window-size = 6 and $\lambda = 0.002$ as the default setting. Our method was compared against four sequence-based methods JSD, VJSD, Consurf, and CRpred, on the dataset Data604. JSD is a sequence conservation based method which uses amino acid position specific frequencies [5]. VJSD takes both stereo-chemical property and residues frequencies into account [19]. Consurf incorporates both sequence conservation information and evolutionary relations among the protein and its homologous sequences [42]. CRpred is an SVM based method which takes five types of attributes into account, including (1) residue type, (2) position specific scoring matrix (PSSM), (3) Shannon entropy computed over the weighted observed percentages (WOP) vector, (4) averaged cumulative hydrophobicity and (5) catalytic resides pairs [37]. Of the four methods used for comparison, JSD, VJSD, and Consurf do not need a training procedure, while CRpred does and therefore it was trained using the same procedure as our method. The optimal parameters of CRpred were obtained from [37], and we tested CRpred with the same ten-fold cross validation procedure as L1pred. The comparison results are shown in Table 1. L1pred shows the best values in terms of both AUPR and AUC, in detail, resulting to AUPR = 0.2198 and AUC = 0.9494. Moreover, L1pred is significantly better than the other four methods (with P-value = $1.24 \times 10^{-6} < 0.05$), according to the ROC significance test. Figure 1 shows the PR curves for all five methods, and the PR curve of L1pred is constantly higher than that of the other PR curves in the whole range of recall rate.

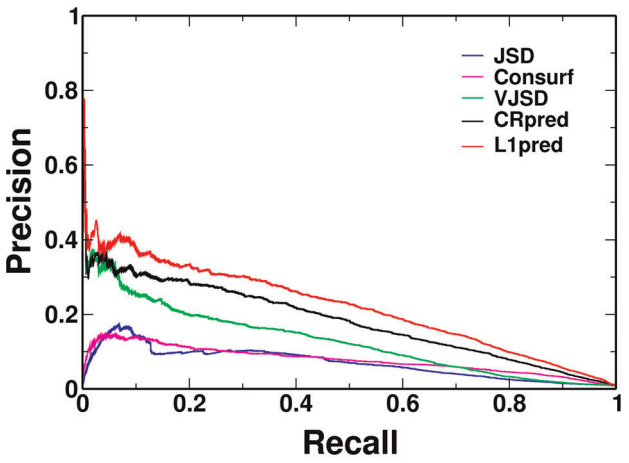

Figure 1. PR curves of five methods on the Data604 dataset.
doi:10.1371/journal.pone.0035666.g001

Table 1. Performance on the dataset Data604.

Method	AUPR	AUC	Recall	Precision
JSD	0.0692	0.8443	0.3299	0.1016
Consurf	0.0778	0.8969	0.3515	0.0944
VJSD	0.1300	0.8700	0.3724	0.1593
CRpred	0.1819	0.9338	0.3805	0.2310
L1pred	0.2198	0.9494	0.3741	0.2752

doi:10.1371/journal.pone.0035666.t001

Results on the Independent Test Dataset Data63

All chosen methods were also compared using the independent test set, Data63, and the results were in broad agreement with what found on the dataset Data604. For L1pred and CRpred, their trained models were generated on the whole Data604 dataset. All results are shown in Table 2, and L1pred shows the best performance. For example, L1pred has the highest values of AUPR and AUC of 0.2636 and 0.9375, respectively. We also tested the statistical significance among different methods in terms of the AUC values. L1pred is significantly better than the other methods; all comparisons showed P-values $< 10^{-10}$, except with CRpred method (P-value = 9.37×10^{-3}), but it is still significant for the cutoff of P-value = 0.05. From the PR curve, shown in Figure 2, one may find that the PR curve of L1pred is notably higher than that of CRpred, the second best method. Especially, if using recall rate = 0.1, the precision of L1pred is more than 60%, while the second best performer is less than 40%. However, all precisions drop fast; at the maximal F-measure point, *i.e.* recall = 0.3571, even the precision of L1pred drops to only 0.3257. These results indicate that L1pred achieves comparable performance on independent dataset with the trained parameters.

CRpred and L1pred have different attribute sets and classifiers. Additional analysis was conducted to figure out which one is essential in prediction. We applied L1-logred classifier to the attribute sets of CRpred method (CRpred-L1) and SVM to attributes of L1pred (L1pred-SVM). All parameters were optimized as the same procedure described in the section of Methods. For the dataset Data604 with ten-fold cross validation, the AUC value of CRpred-L1 is 0.9341, which is approximately equal to that of CRpred, 0.9338. L1pred and L1pred-SVM also have close AUC values on the dataset Data604; they are 0.9494 and 0.9480, respectively. The situation for the dataset Data63 is similar as well. These results indicate that the combination of those eight attributes used by L1pred plays important role in the improvement of prediction performance.

Moreover, L1pred is more efficient than other machine learning methods, *e.g.* the SVM-based CRpred method, because L1-logred

Table 2. Performance on the dataset Data63.

Method	AUPR	AUC	Recall	Precision
JSD	0.0759	0.8410	0.4160	0.1061
Consurf	0.1019	0.8876	0.2017	0.1644
VJSD	0.1520	0.8599	0.3109	0.2349
CRpred	0.1809	0.9201	0.4244	0.2446
L1pred	0.2636	0.9375	0.3571	0.3257

doi:10.1371/journal.pone.0035666.t002

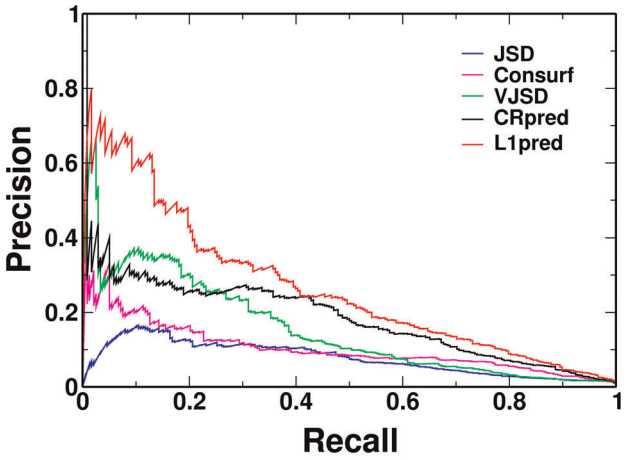

Figure 2. PR curves of five methods on the Data63 dataset.
doi:10.1371/journal.pone.0035666.g002

is a fast classifier. Table 3 shows the result of the comparison between L1pred and CRpred in terms of computing time for training and testing. L1pred is about 40 times faster than CRpred in both training and predicting.

Results on the Dataset EF-family

L1pred and CRpred are applied on the dataset EF-family with the same ten-fold cross validation procedure. For CRpred, the encoded feature vector of each protein was directly downloaded from their web site [37]. This dataset has been used by Youn *et al.* to test their structure-based method [36], which integrated several different types of attributes, including structural conservation, B-factor, solvent accessibility, and sequence conservation *etc*.

From Table 4, which shows all results on the dataset EF-family, one may find that the overall results are similar to that on both Data604 and Data63. Specifically, the values of AUPR and AUC of L1pred, 0.2589 and 0.9372, are higher than those of CRpred. The difference between L1pred and CRpred is significant for ROC, with a P-value of 1.61×10^{-11}. Moreover, L1pred is also slightly better than Youn's method in terms of AUC. The result of Youn's method on the dataset of EF-family was obtained directly from their publication [36].

Importance of Different Features

To understand which attributes of all eight different scores play more important roles, we removed them one by one and repeated the same training and validation procedure on the dataset Data604. The results are shown in Table 5. One may find that the omission of any score leads to some changes in performance, but none was significant. The largest drop occurred when the Consurf score was turned off. We therefore concluded that all eight attributes are almost equally important for L1pred, but the Consurf score is slightly more important than all others.

We also extracted the weight vector of the trained model on the whole Data604 dataset. The top 15 weighted bits are shown in Figure 3 in which, for example, SS-4-E denotes the SS attribute of the beta strand at the 4th position on the N-terminal side of the central bit in a sliding window. The similar notations are applied for the other features, and $i>0$ represents positions towards C-terminal, $i=0$ represents the central residue and $i<0$ represents positions towards N-terminal. We found that VJSD+0 has the largest weight, which means the stereo-chemical characteristics are correctly reflected by this scoring function, and the majority of catalytic residues can be distinguished by this feature. In addition, being a Cys residue (RT-Cys) and/or a charged/polar residue (OP-Polar, OP-Charged) are important features for catalytic sites, which agrees with the statistical results [43]. In the trained model, the Consurf score of position 0 is also important for catalytic residues prediction as ranked on the third position. Assigning a large weight to ACH-Win17 indicates that the mean hydrophobicity of 16 residues around the catalytic residues plays an important role for catalytic functions. These results suggest that L1pred can extract the most useful chemical/physical characteristics of catalytic residues by the training procedure.

Table 3. Computing time of L1pred and CRpred methods.

Method	AUPR	AUC	Recall	Precision
JSD	0.0759	0.8410	0.4160	0.1061
Consurf	0.1019	0.8876	0.2017	0.1644
VJSD	0.1520	0.8599	0.3109	0.2349
CRpred	0.1809	0.9201	0.4244	0.2446
L1pred	0.2636	0.9375	0.3571	0.3257

doi:10.1371/journal.pone.0035666.t003

Table 4. Performance on the dataset EF-family.

Method	AUPR	AUC	Recall	Precision
JSD	0.0841	0.8543	0.0886	0.5522
Consurf	0.0969	0.8767	0.1229	0.3048
VJSD	0.1695	0.8873	0.2333	0.2756
CRpred	0.2256	0.9118	0.2853	0.3838
Youn	N/A	0.9298	0.5702	0.1851
L1pred	0.2589	0.9372	0.4478	0.2862

doi:10.1371/journal.pone.0035666.t004

Case Studies

We randomly selected two enzymes from our datasets as examples to show the prediction performance of L1pred; they are a dehydrogenase (PDB ID: 1A05 chain A) and an asparaginase

Table 5. Performance of L1pred by removing attributes one by one.

Method	AUPR	AUC	Recall	Precision
no-Consurf	0.1688	0.9282	0.3854	0.2125
no-SS	0.2119	0.9467	0.4559	0.2440
no-RT	0.2128	0.9492	0.4370	0.2455
no-ACH	0.2129	0.9486	0.4736	0.2313
no-VJSD	0.2140	0.9488	0.4392	0.2466
no-JSD	0.2167	0.9492	0.4623	0.2422
no-ASA	0.2175	0.9494	0.3947	0.2640
no-OP	0.2184	0.9487	0.4128	0.2607
L1pred	0.2198	0.9494	0.3741	0.2752

doi:10.1371/journal.pone.0035666.t005

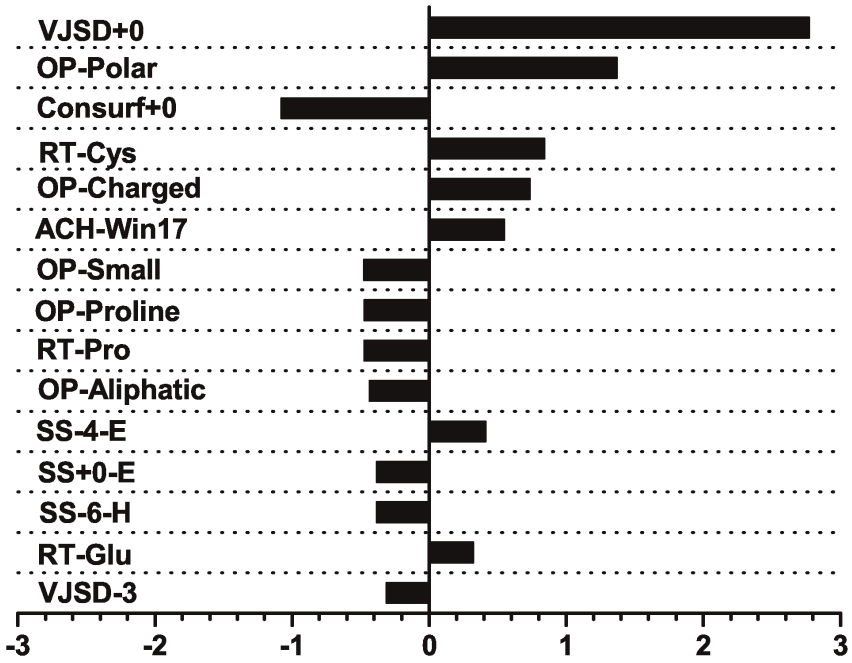

Figure 3. Weights of the top fifteen features on the Data604 dataset.
doi:10.1371/journal.pone.0035666.g003

(PDB ID: 3ECA chain A). There are three catalytic residues for the dehydrogenase (140Y, 190L, and 222D) and five for the asparaginase (12T, 25Y, 89T, 90D, and 162L). Prediction results of L1pred are shown in Figure 4. For each enzyme, true catalytic residues and 10 top-ranked residues are shown in colors; correctly predicted catalytic residues are shown in red, missed catalytic residues (false negative) in blue, and the resides predicted by L1pred but not true catalytic residues (false positive) in green. Two out of three catalytic residues were correctly predicted for the dehydrogenase and four out of five for the asparaginase. Both cases indicate that L1pred can discover more than 60% catalytic residues with recall = 4%, as the lengths of those enzymes are both more than 300 amino acids.

Discussion

We applied the L1-logreg classifier with eight attributes to predict enzyme catalytic residues. The attributes, VJSD, overlapping properties, and Consurf score, are newly introduced to the solution of catalytic residue prediction. With the ten-fold cross validation on the dataset Data604 and directly application on the independent test set Data63, L1pred showed the best performance among chosen algorithms. The AUC values of L1pred on the dataset Data604 and Data63 are 0.9494 and 0.9375, respectively, which are significantly higher than other prediction methods (P-value$<$0.05). The test on the EF-family dataset confirms that this method performs better than existing methods, including the structure-based one. In all eight attributes, Consurf, SS, RT, and averaged cumulative hydrophobicity play slightly more important

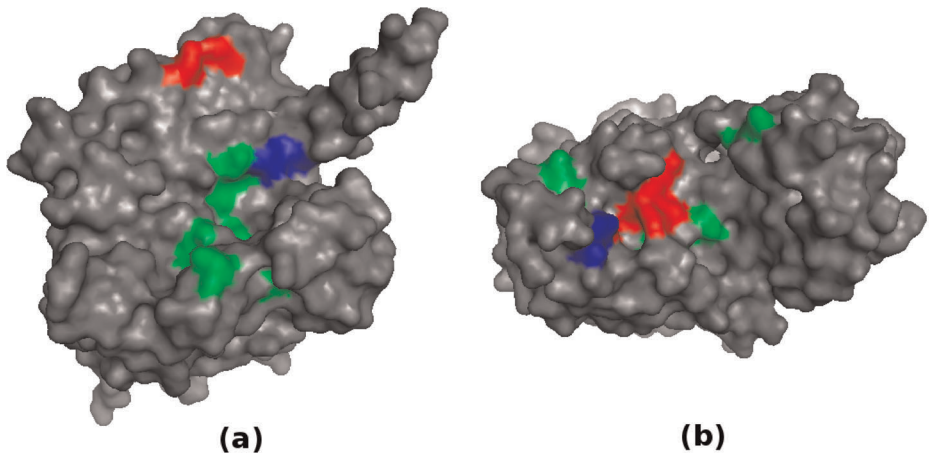

Figure 4. Prediction results of L1pred on a dehydrogenase (a) and an asparaginase (b) Red: true positive, blue: false negative, and green: false positive.
doi:10.1371/journal.pone.0035666.g004

roles than the other attributes. The scoring functions of Consurf and VJSD used in this manuscript can be combined with structural information to improve catalytic residue prediction. Further analysis indicates that the improvement made by L1pred is mainly due to the combination of informative attributes, instead of the classifier. L1-logreg classifier is not necessary to have better performance in catalytic residue prediction than SVM, but it is efficient and hence competent for genome-wide analyses, where speed is an issue. In the future, we will test additional scoring functions to further improve the prediction performance, and extend the platform developed for this project to other applications, such as protein phosphorylation site prediction.

Materials and Methods

Datasets

We collected our data from two sources: the datasets created by Zhang *et al.* [37] and the Catalytic Site Atlas (CSA) dataset [44]. By mixing the CSA and the eight datasets from Zhang *et al.* [37], (namely, EF-family, EF-fold, EF-superfamily, HA-superfamily, NN, PC, T-124, and T-37), we generated two new datasets. Since all enzymes in our datasets have structures in PDB, we first compared their sequences with the sequences of the structures in PDB [45]. If two sequences are not identical, this enzyme was discarded. For the remaining protein sequences, we clustered them using Blastclust [46] with sequence identity 30% and coverage 60%. A total of 667 clusters were returned, 604 of which have single members and 63 have multiple ones. Those 604 chains with sequence similarity lower than 30.0% to the other chains were selected as a dataset and named Data604. For the other 63 clusters, we randomly picked one protein sequence from each cluster and gathered them as another dataset called Data63. The Data63 is used as an independent test dataset in the study. For both datasets, we randomly selected six non-catalytic residues for one catalytic residue in each sequence. To further compare L1pred and CRpred directly, all chosen methods are compared on the EF-family dataset from [37]. Proteins in this data set that are not the same as or part of the corresponding sequences in PDB were discarded, and 347 chains were left.

Classifier Feature Vectors

Here, we first describe construction of feature vectors. For a given amino acid residue, we collect a sub-sequence with all residues adjacent to it by a certain window size, e.g. 4, which means the total length of this sub-sequence is 4+1+4 = 9. For this sub-sequence, we encode it with a multidimensional vector based on eight sequence-based attributes. The L1-logreg classifier is then applied to these vectors to train a model and then predict catalytic residues. The eight attributes we use are residue type (RT), overlapping properties (OP), averaged cumulative hydrophobicity (ACH), Jensen-Shannon divergence conservation score (JSD), the combination of relative entropy of Venn diagram and JSD conservation score (VJSD), predicted protein secondary structure (SS), predicted solvent accessible surface area (ASA), and Consurf score. In the following, we describe each attributes in details.

Residue Type (RT). RT is a commonly used attribute for protein-sequence-based machine learning methods. Each amino acid is encoded by a 20-bit binary vector where the dimension of the corresponding amino acids is set to 1 and others are 0, *i.e.*, A (10000000000000000000), … V (00000000000000000001). The order of amino acids in this manuscript is A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V.

Overlapping Properties (OP). Several previous studies suggested that the Taylor's overlapping properties are useful for catalytic residues prediction [8,19]. These properties are: Polar [NQSDECTKRHYW], Positive [KHR], Negative [DE], Charged [KHRDE], Hydrophobic [AGCTIVLKHFYWM], Aliphatic [IVL], Aromatic [FYWH], Small [PNDTCAGSV], Tiny [ASGC] and Proline [P] [47]. Residues are encoded using 10-bit vectors where the dimensions of the corresponding properties are set to 1 and remaining positions are 0, *i.e.*, A (0000100010), … V (0000110100).

Averaged Cumulative Hydrophobicity (ACH). ACH has been demonstrated to be an important attribute for catalytic residues [37]. The attribute is extracted by computing the average of the cumulative hydrophobicity indices over a window with size varying as 3, 5, 7, …, 21. As a result, ten ACH scores are extracted. Hydrophobicity index proposed by Sweet and Eisenberg [48] is used in the paper. If the central residue is at the sequence termini, we use 0s to fill in the blanks.

Jensen-Shannon divergence (JSD) scores. WOP is another important information source extracted by PSI-BLAST [46]. The WOP vector for a position represents the position-specific distribution of 20 amino acids. It has been used to calculate sequence conservation in several previous works [19,37] and is used as the source of amino acid position-specific distribution in the study. The JSD score of a residue S is computed as:

$$\mathrm{JSD}_S = \frac{1}{2}\sum_{i=1}^{20} p_S(i)\log\frac{p_S(i)}{\frac{1}{2}p_0(i)+\frac{1}{2}p_S(i)}$$

$$+\frac{1}{2}\sum_{i=1}^{20} p_0(i)\log\frac{p_0(i)}{\frac{1}{2}p_S(i)+\frac{1}{2}p_0(i)},$$

where $p_s(i)=a_i/\sum_1^{20} a_i$, a_i is the ith WOP value at the site ($i=1,\cdots,20$) and p_0 is the BLOSUM62 amino acid background distribution.

Combination of relative entropy of Venn diagram and JSD (VJSD). The relative entropy of Venn diagram (RVD) score is based on Taylor's Vine diagram of amino acids as shown above in the overlapping properties [19]. Calculating RVD scores needs the WOP matrix from PSI-BLAST as well. The RVD score of the residue on site S is defined as:

$$\mathrm{RVD}_S = \sum_{i=1}^{10} p_S(i)\ln\frac{p_S(i)}{p_0(i)},$$

where $p_S(i)$ is the fractional WOP values of all residues with the same property i in the site S, *i.e.* $p_S(i)=\sum_{k\in i} a_k/\sum_{j=1}^{20} a_j$, a_j is the jth WOP value, and $p_0(i)$ is the fractional BLOSUM62 value of the same class i for the background distribution.

Taylor's Vine diagram can not distinguish residues, such as TYR and TRP, GLY and ALA, ILE and LEU. But methods which based on residue frequencies can discriminate them naturally. Therefore, RVD is combined with JSD, which is based on residue frequencies, to overcome the weakness. The combined score of a residue S is given by:

$$\mathrm{VJSD}_S = \sqrt{nRVD_S^2 + nJSD_S^2},$$

where the $nRVD_S$ and $nJSD_S$ are the normalized RVD and JSD scores of the site.

Predicted protein secondary structure (SS). Previous study suggested that more than 50% catalytic residues occur in coil regions of proteins [43]. Therefore, the protein secondary structure deserves to be considered as an attribute in catalytic residue prediction. The most accurate way to obtain the information of secondary structure would be from the 3D structures of proteins, but for a given protein sequence, currently, we can only predict the secondary structures. In this manuscript, the SS attribute of each residue has three bits to show the possibility scores of three types of secondary structures (H, E, and C), which is predicted by PSIPRED [49].

Predicted accessible surface area (ASA). All catalytic residues are on the surface of enzyme proteins, and hence, large solvent accessibility is also an important feature for the catalytic residues. To improve the prediction accuracy, we combined ASA into our frame as well. For the same reason as for the case of SS, the ASA attribute is also predicted with protein sequences. In this study, we used RVP-net [50] to predict the solvent accessible surface area for each residue for a give protein sequence. Each residue has a value of 0 or 1 for the ASA attribute.

Consurf score. The Consurf method is based on evolutionary relations among proteins represented by phylogenetic trees [42]. It was used to predict functional sites of proteins by estimating the degree of sequence conservation among their homologous sequences [51,52]. Consurf scores of all proteins were obtained from the web server http://consurfdb.tau.ac.il.

When appling the sliding window strategy to a given protein sequence with the above eight scores, we devised a few modifications to circumvent issues. If a residue on a sequence terminus is the central bit of a sliding window, we use 0s to fill in blanks on one side of the window. For attributes RT, OP, and ACH, we just applied them to the central bit of a sliding window, making them independent of the size of the windows.

L1-logreg classifier. We use the L1-logreg classifier to score and classify all data vectors, and hence, predict catalytic residues. The classifier is a large-scale solver for L1-regularized logistic regression problems [41], which has been proven to yield models better than those based on unregularized estimations [53–55]. For the given data vectors, $x \in \mathrm{R}^n$, to be classified, the logistic model calculates the conditional probability of $s \in \{-1,1\}$,

$$\mathrm{P}(s|x) = \frac{\exp(s(w^T x + v))}{1 + \exp(s(w^T x + v))}.$$

The model has parameters $w \in \mathrm{R}^n$ (the weight vector) and $v \in \mathrm{R}$ (the intercept); $w^T x + v = 0$ defines the neutral hyper-plane in the data vector space. The classifier locates the optimal model by maximizing the likelihood estimation from the observed examples, *i.e.* minimizing the average logistic loss:

$$\min \quad (1/m)\sum_{i=1}^{m} \log(1 + \exp(-s_i(x_i^T w + v))) + \lambda \sum_{i=1}^{n} |w_i|,$$

where $\lambda > 0$ is the regularization parameter, which is used to balance the average logistic loss and the size of the weight vector. More details on the L1-logreg classifier can be found in reference [41]. We used the software package of L1-logreg classifier as implemented by [41] and available at

http://www.stanford.edu/~boyd/l1_logreg/.

Training and Testing Procedure

The parameter λ of L1-logreg and the window size were optimized on the dataset Data604 with a ten-fold cross validation. The optimal set of window size and λ that gives rise to the highest AUPR values, were obtained by a grid search in the interval of [0.001, 0.02] with a step of 0.001 for λ and from 0 to 10 for the window size. For each duplet, the ten-fold cross validation procedure was used to test the performance. Once obtaining the optimal values for window size and λ, we trained the model on the whole set of Data604 for the test and real prediction. To determine the optimal point of precision and recall rate on the ROC curve, we used the F-measure that is defined in the following equation:

$$\mathrm{F} = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}},$$

where P and R are Precision and Recall rate, respectively. Please see the section of Evaluation for definitions. We took trained parameters that perform with the maximal F-measure point [56], which is the balance point of sensitivity and specificity.

Evaluation

To evaluate the performance of our method, we used Precision (P), Recall (R), False Positive Rate (FPR). They are defined by the following equations:

$$\mathrm{P} = \frac{TP}{TP + FP},$$

$$\mathrm{R} = \frac{TP}{TP + FN},$$

$$\mathrm{FPR} = \frac{FP}{TN + FP},$$

where TP, TN, FP and, FN are the true positive, true negative, false positive, and false negative rate, respectively. To compare among different algorithms, all P, R, and FPR are calculated at the point with the maximal F-measure. The area under the Precision-Recall (PR) curve (AUPR) is also used to evaluate the performances of all methods. A receiver operating characteristic (ROC) curve represents a dependency of sensitivity and (1-specificity). To obtain the ROC curve, all sites in a dataset are sorted by their scores, and we increase the number of predicted sites in steps of one site each time. In addition, the online tool, StAR, is used to test the statistical significance between AUC values [57].

Author Contributions

Conceived and designed the experiments: YD CZ. Performed the experiments: YD. Analyzed the data: YD. Wrote the paper: CZ YD JW JY.

References

1. Mirny L, Shakhnovich E (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 291: 177–196.
2. Valdar W (2002) Scoring residue conservation. Proteins 48: 227–241.
3. Pei J, Grishin N (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17: 700–712.
4. Wang K, Samudrala R (2006) Incorporating background frequency improves entropy-based residue conservation measures. BMC Bioinformatics 7: 385.

5. Capra J, Singh S (2007) Predicting functionally important residues from sequence conservation. Bioinformatics 23: 1875–1882.
6. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 13: 190–202.
7. Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, et al. (2004) Conseq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics 20(8): 1322–1324.
8. Dou YC, Zheng XQ, Wang J (2009) Several appropriate background distributions for entropy-based protein sequence conservation measures. J Theor Biol 262(2): 317–322.
9. Mayrose I, Graur D, Ben-Tal N, Pupko T (2004) Comparison of sitespecific rate-inference methods for protein sequences: Empirical bayesian methods are superior. Mol Biol and Evol 21: 1781–1791.
10. Sterner B, Singh R, Berger B (2007) Predicting and annotating catalytic residues: an information theoretic approach. J Comput Biol 14: 1058–1073.
11. Zhang SW, Zhang YL, Pan Q, Cheng YM, Chou KC (2008) Estimating residue evolutionary conservation by introducing von neumann entropy and a novel gap-treating approach. Amino Acids 35: 495–501.
12. Sankararaman S, Sjölander K (2008) Intrepidinformation-theoretic tree traversal for protein functional site identification. Bioinformatics 24: 2445C2452.
13. La D, Sutch B, Livesay DR (2005) Predicting protein functional sites with phylogenetic motifs. Proteins 58: 309–320.
14. Bahadur Dukka KC, Livesay DR (2008) Improving position-specific predictions of protein functional sites using phylogenetic motifs. Bioinformatics 24: 2308–2316.
15. Ye K, Vriend G, IJzerman AP (2008) Tracing evolutionary pressure. Bioinformatics 24: 908–915.
16. Mihalek I, Reš I, Lichtarge O (2004) A family of evolution-entropy hybrid methods for ranking residues by importance. J Mol Biol 336: 1265–1282.
17. Fischer JD, Mayer CE, Söding J (2008) Prediction of protein functional residues from sequence by probability density estimation. Bioinformatics 24: 613–620.
18. Dou YC, Zheng XQ, Wang J (2009) Prediction of catalytic residues using the variation of stereochemical properties. Protein J 28: 29–33.
19. Dou YC, Zheng XQ, Yang JL, Wang J (2010) Prediction of catalytic residues based on an overlapping amino acid classification. Amino Acids 39: 1353–1361.
20. Liu XS, Guo WL (2008) Robustness of the residue conservation score reecting both frequencies and physicochemistries. Amino Acids 34: 643–652.
21. Williamson RM (1995) Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters. J Theor Biol 24: 908–915.
22. del Sol Mesa A, Pazos F, Valencia A (2003) Automatic methods for predicting functionally important residues. J Mol Biol 326: 1289–1302.
23. Innis CA, Anand AP, Sowdhamini R (2003) Prediction of functional sites in proteins using conserved functional group analysis. J Mol Biol 337: 053–1068.
24. Panchenko AR, Kondrashov F, Bryant S (2003) Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci 13: 884–892.
25. Tang YR, Sheng ZY, Chen YZ, Zhang ZD (2008) An improved prediction of catalytic residues in enzyme structures. Protein Eng Des Sel 21: 295–302.
26. Alterovitz R, Arvey A, Sankararaman S, Dallett C, Freund Y, et al. (2009) Resboost: characterizing and predicting catalytic residues in enzymes. BMC Bioinformatics 10: 197.
27. Chea E, Livesay DR (2007) How accurate and statistically robust are catalytic site predictions based on closeness centrality? BMC Bioinformatics 8: 153.
28. Tong W, Wei Y, Murga LF, Ondrechen MJ, Williams RJ (2009) Partial order optimum likelihood (pool): Maximum likelihood prediction of protein active site residues using 3d structure and sequence properties. PLoS Comput Biol 5(1): e1000266.
29. Gong S, Blundell TL (2008) Discarding functional residues from the substitution table improves predictions of active sites within three-dimensional structures. PLoS Comput Biol 4(10): e1000179.
30. Marino Buslje C, Teppa E, Di Domenico T, Delfino JM, Nielsen M (2010) Networks of high mutual information define the structural proximity of catalytic sites: Implications for catalytic residue identification. PLoS Comput Biol 6(11): e1000978.
31. Lopez G, A V, Tress M (2007) firestarprediction of functionally important residues using structural templates and alignment reliability. Nucleic Acids Res 35: W573–W577.
32. Brylinski M, Prymula K, Jurkowski W, Kochanczyk M, Stawowczyk E, et al. (2007) Prediction of functional sites based on the fuzzy oil drop model. PLoS Comput Biol 3(5): e94.
33. Gutteridge A, Bartlett GJ, Thornton JM (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. J Mol Biol 303: 719–734.
34. Pande S, Raheja A, Livesay DR (2007) Prediction of enzyme catalytic sites from sequence using neural networks. IEEE symposium on CIBCB 07: 247–253.
35. Petrova N, Wu C (2006) Prediction of catalytic residues using support vector machines with selected protein sequence and structural properties. BMC Bioinformatics 7: 312.
36. Youn E (2007) Evaluation of features for catalytic residue prediction in novel folds. Protein Sci 16: 216–226.
37. Zhang T, Zhang H, Chen K, Shen SY, Ruan JS, et al. (2008) Accutate sequence-based prediction of catalytic residues. Bioinformatics 24: 2329–2338.
38. Sankararaman S, Sha F, Kirsch JF, Jordan MI, Sjölander K (2010) Active site prediction using evolutionary and structural information. Bioinformatics 5: 617–624.
39. Cilia E, Passerini A (2010) Automatic prediction of catalytic residues by modeling residue structural neighborhood. BMC Bioinformatics 11: 115.
40. Kato T, Nagano N (2010) Metric learning for enzyme active-site search. Bioinformatics 26: 2698–2704.
41. Koh K, Kim SJ, Boyd S (2007) An interior-point method for large-scale l1-regularized logistic regression. J Mach Learn Res 8: 1519–1555.
42. Armon A, Graur D, Ben-Tal N (2001) Consurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 307: 447–463.
43. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM (2002) Analysis of catalytic residues in enzyme active sites. J Mol Biol 324: 105–121.
44. Porter C, Bartlett G, Thornton J (2003) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32: D129–D133.
45. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucleic Acids Res 28: 235–242.
46. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25;17): 3398–3402.
47. Taylor W (1986) The classification of amino acid conservation. J Theor Biol 119: 205–218.
48. Sweet RM, Eisenberg D (1983) Correlation of sequence hydrophobicities measures similarity in three dimensional protein structure. J Mol Biol 171: 479–488.
49. McGuffin LJ, Bryson K, Jones DT (2000) The psipred protein structure prediction server. Bioinformatics 16: 404–405.
50. Ahmad S, Gromiha MM, Sarai A (2003) RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. Bioinformatics 19(14): 1849–1851.
51. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, et al. (2008) Consurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 19: 163–164.
52. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, et al. (2005) Consurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 33: W299–W302.
53. Greenshtein E, Ritov Y (2004) Persistence in high-dimensional predictor selection and the virtue of overparametrization. Bernoulli 10: 971–988.
54. Zhao P, Yu B (2006) On model selection consistency of lasso. J Mach Learn Res 7: 2541–2563.
55. van de Geer SA (2008) High-dimensional generalized linear models and the lasso. Ann Stat 36: 614–645.
56. Liu ZP, Wu LY, Wang Y, Zhang XS, Chen LN (2010) Prediction of protein-rna binding sites by a random forest method with combined features. Bioinformatics 26: 1616–1622.
57. Vergara IA, Norambuena T, Ferrada E, Slater AW, Melo F (2008) StAR: a simple tool for the statistical comparison of roc curves. BMC Bioinformatics 9: 265.